CPSC 4300/6300 Applied Data Science

Lab Session 1 – Introduction to R

1 Installation

Install the most recent versions of R and RStudio for your operating system from https://www.r-project.org/ and https://www.rstudio.com/.

2 Basic commands

This is a comment block:

```
# Author: Alex Herzog
# Date: August 24, 2018
# Purpose: Introduction to R
```

Assigning data to an object:

```
x <- c(4, 8, 15, 16, 23, 42)
```

Basic summary statistics:

```
length(x)
mean(x)
min(x)
max(x)
sd(x)
summary(x)
```

Create more objects:

```
y <- c(8, 27, 34, 4, 19, 10)
z <- c(1, 2)
```

Combine objects:

```
a <- x + y
b <- x + z
```

Check what is currently in the workspace:

```
ls()
```

Remove objects:

```
rm(x)
rm(a, b)
```

Help files:

```
?matrix
help(matrix)
```

3 Matrices and indexing

Create a matrix:

```
x <- matrix(data=c(1, 2, 3, 4), nrow=2, ncol=2)
```

Create the same matrix but with shorter notation:

```
x <- matrix(c(1, 2, 3, 4), 2, 2)
```

Simple operations on the elements of a matrix:

```
x^2
sqrt(x)
```

Indexing:

```
x[1,]
x[,1]
x[1,2]
```

4 Graphics

Generate two vectors with random numbers:

```
set.seed(42)
x <- rnorm(50)
y <- x + rnorm(50, mean=50 ,sd=.1)
```

Scatterplot of two vectors:

```
plot(x, y)
plot(x ,y , xlab="Advertising budget", ylab="Sales", main="Simulated data")
```

Adding a regression line:

```
abline(lm(y ~ x), col="red")
```

5 Loading data

Read a data file:

```
Auto <- read.table("http://www-bcf.usc.edu/~gareth/ISL/Auto.data")
```

Look at the content of the file. What do you observe?

```
dim(Auto)
head(Auto)
View(Auto)
```

Read the data file again:

```
Auto <- read.table("http://www-bcf.usc.edu/~gareth/ISL/Auto.data", header=TRUE,
   na.strings="?")
View(Auto)
```

Check the column headers:

```
names(Auto)
```

Remove missing values:

```
Auto <- na.omit(Auto)
```

Indexing a data frame:

```
Auto$year
table(Auto$year)
```

6 Factor objects

Look at the column `cylinders`:

```
summary(Auto$cylinders)
table(Auto$cylinders)
plot(Auto$cylinders)
```

Convert column `cylinders` to a factor variable:

```
Auto$cylinders <- as.factor(Auto$cylinders)
plot(Auto$cylinders)
```

7 Exercise

This exercise relates to the `College` data set, which can be found in the file http://www-bcf.usc.edu/~gareth/ISL/College.csv. It contains a number of variables for 777 different universities and colleges in the US. The variables are

- `Private`: Public/private indicator
- `Apps`: Number of applications received
- `Accept`: Number of applicants accepted
- `Enroll`: Number of new students enrolled
- `Top10perc`: New students from top 10 % of high school class
- `Top25perc`: New students from top 25 % of high school class
- `F.Undergrad`: Number of full-time undergraduates
- `P.Undergrad`: Number of part-time undergraduates
- `Outstate`: Out-of-state tuition
- `Room.Board`: Room and board costs
- `Books`: Estimated book costs
- `Personal`: Estimated personal spending
- `PhD`: Percent of faculty with Ph.D.'s

- `Terminal`: Percent of faculty with terminal degree
- `S.F.Ratio`: Student/faculty ratio
- `perc.alumni`: Percent of alumni who donate
- `Expend`: Instructional expenditure per student
- `Grad.Rate`: Graduation rate

1. Use the `read.csv()` function to read the data into R. Call the loaded data `college`.

2. Look at the data using the `View()` function. You should notice that the first column is just the name of each university. We don't really want R to treat this as data. However, it may be handy to have these names for later. Try the following commands:

```
rownames(college) <- college[ ,1]
View(college)
```

   Now you should see that the first data column is `Private`. Note that another column labeled `row.names` now appears before the `Private` column. However, this is not a data column but rather the name that R is giving to each row.

3. Use the `summary()` function to produce a numerical summary of the variables in the data set.

4. Use the `pairs()` function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix `A` using `A[,1:10]`.

5. Use the `plot()` function to produce side-by-side boxplots of Outstate versus Private.

6. Create a new qualitative variable, called `Elite`, by binning the `Top10perc` variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.

```
college$Elite <- rep("No", nrow(college))
college$Elite[college$Top10perc > 50] <- "Yes"
college$Elite <- as.factor(college$Elite)
```

   Use the `summary()` function to see how many elite universities there are. Now use the `plot()` function to produce side-by-side boxplots of `Outstate` versus `Elite`.

7. Use the `hist()` function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command `par(mfrow=c(2,2))` useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.

8. Continue exploring the data, and provide a brief summary of what you discover.

8 Questions

1. What is the university with the most students in the top 10% of class?
2. What university has the smallest acceptance rate?
3. What university has the most liberal acceptance rate?
4. What is the correlation between out-of-state tuition and graduation rate?